A Robot in a Box

Artur M. Arsenio

Electrical Engineering and Computer Science Department
Artificial Intelligence Lab - Massachusetts Institute of Technology
545 Technology Square, Room NE43-936, MA 02139
arsenio@ai.mit.edu

Abstract

Pervarsive robotics will require small, light and cheap robots that exhibit complex behaviors. Design flexibility, a modular software/hardware architecture and a large repertoire of functionalities are all requirements for such a comprehensive vision system. These demands led to the development of the M2-M4 Macaco project - a robotic active vision head. Macaco is a portable system, capable of emulating the head of different creatures both aesthetically and functionally. It integrates mechanisms for social interactions, autonomous navigation and object analysis.

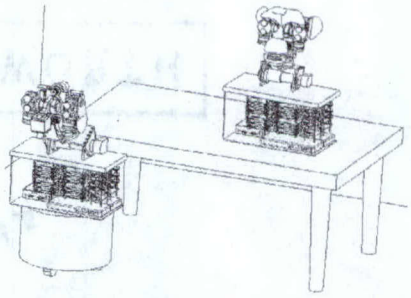

Figure 1: *A portable robotic head, aesthetically flexible and that demonstrates complex behaviors mounted on a mobile platform or as a stand-alone system.*

1 Introduction

The M2-M4 Macaco, a robotic active, modular and compact vision system, was designed to fit different mobile robot platforms or act as a stand-alone system, as illustrated by Figure 1. The creature repertoire of tasks extends the capabilities of a social agent to more functional tasks. An important goal of the mechanical and electronic designs was the portability of both the mechanical device and its brain. A simple communications interface enables operation onboard mobile platforms, turning the robot into an autonomous social machine. Macaco characteristics make it a portable, fully operational robotic head whenever not assembled to a mobile platform, able to act as a social agent.

Research robots are emprisioned most of the time in lab facilities where they are developed, most often operating just for demonstration goals. We expect in a near future to have both the robotic head and its brain physically present at exhibitions/seminars, interacting socially. This new approach with complex, portable research robots will lead to comercial applications and increasing sinergy among roboticists. Eventualy, pervarsive robotics - robots present everywhere to perform a variety of tasks - will be possible as smaller, lighter and cheaper robots become available.

Previous research work on sociable robots has been mostly applied to slow or stationary unintegrated systems with few motility requirements [2]. Other approaches for mobile robots focused on navigation capabilities or industrial applications such as environment reconstruction/map building. Scalability has been an important issue for such mobile systems, and few emphasis has been placed on social mobile robots.

Security is one possible operational scenario for this active head. For this class of applications, it is crutial for the robot to search for people (behavioral modes presented in Section 4) - or faces, and to further recognize them, as described in Section 6.1. In addition, human gaze direction might reveal security threats, and thus Section 6.2 introduces a head gaze detection algorithm. Probable targets for such gazings are other people and mostly important, explosives and/or guns. Therefore, salient objects situated in the world are processed for 3D information extraction (Section 5.2) and texture/color analysis (Section 5.1). Current work is also underway for object and scene recognition from contextual cues (Section 8).

Another scenario includes search and rescue missions by a mobile robot, which requires additional

navigation capabilities in rough terrain (introduced in Section 7). Finally, real world applications are often characterized by strong light variations or the absence of light. This is taken into account through termal image processing for people detection (Section 3.3) and for night navigation (Section 7.1).

2 The Active Vision Heads

A robotic mechanism designed to act as an intelligent creature should explore the mechanisms which evolution provided these creatures with, and incorporate them on the design. Hence, this robotic head was designed to resemble a biological creature, exploiting several features of an evolutionary design, but adding others (such as a thermal camera for human detection and night vision) for improved performance.

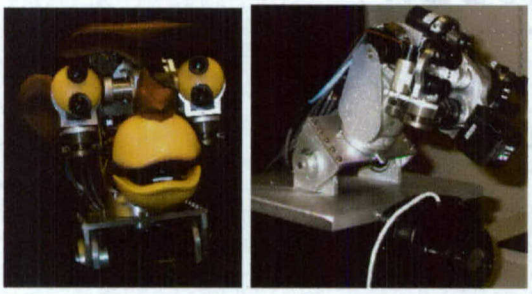

Figure 2: *(left) M2 Macaco, a biological inspired robotic head designed to ressemble a primate head. Aesthetic parts manufactured using the Stereolithography process (right) The M4 Macaco robot, designed to ressemble a dog's head.*

The construction of both dog like and primate like robotic bodies at the MIT LegLab required canine/primate like heads to fit these bodies. One goal of the M2-M4 Macaco project was, therefore, the creation of a flexible robotic head that could match both these quadruped (M4) and biped (M2) robots. The replacement of a few number of M2-M4 Macaco aesthetic components allows for this metamorphosis (Figure 2), while all the remaining mechanical components are functional. The robotic head is easily assembled to a mobile platform, being the communications carried out through a serial port.

The weight of the head, including motors, gears and gyro, is $\sim 3.55lbs$ ($\sim 1.6Kg$). Its height is approximately $7.5in$, and the distance between the eyes is $\sim 3.5in$. The head has seven degrees of freedom (four in the neck, three for the eyes). It also includes two eyes, and two cheap CMOS miniature color board cameras (specially well-suited for dynamic scenes) at each eye - one for the foveal vision, of higher resolution ($31^o \times 23^o$), and the other for a wide field of view ($110^o \times 83^o$). The eyes have a common actuated revolute tilt joint and independent pan actuated joints. An inertial sensor was placed in the nose of the M4 head, and inside of the upper jaw for the M2 head. All motors are controlled by small, compact and light JRKerr controllers, with amplifiers onboard. Motor commands are sent through two serial ports. An additional motor moves M2-Macaco's lower jaw, for future applications in speech synthesis and recognition.

The hardware consists of two PC104+ boards, AMD K6-II at 400MHz processor based, and seven small CPU boards with Pentium III at 800MHz, all modules connected by an Ethernet network. A total of nine framegrabbers are being used for video acquisition, being each camera connected to several grabbers to reduce network latencies. The network server is connected to a notebook hard drive. OS-QNX runs on all processors, featuring high performance and real time transparent communications.

Currently, we are further reducing the size and weight of the overall system, placing all electronics in a small, portable box. We are as well upgrading the hardware to nine PC104+ boards 700MHz processor based. All the hardware, together with the internet hubs, video amplifiers, the intersense gyro control box and the motor controllers will fit into a box. The power electronics for the hardware, cameras and motors fits into a separate box. Not only is this mechanism small and light-weight (designed for portability), but it is also inexpensive ($\sim$ 20.000 USD including mechanical construction and hardware electronics).

The robotic head was also assembled to a *Magellan* mobile platform from *iRobot* (see Figure 3).

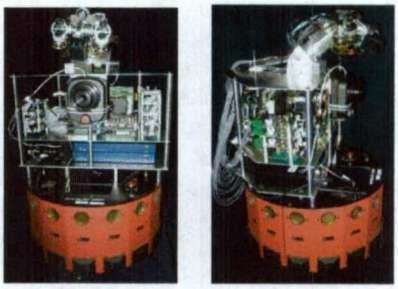

Figure 3: *The M4-Macaco robotic head and processing hardware on top of a Magellan mobile platform.*

3 Perceptual Modules

3.1 Visual Attention System

A convincing intelligent robot is greatly gauged by its putative social skills, together with its navigational

ability in the surrounding environment. A logpolar attentional system was developed to select relevant information from the cameras output, and to combine it in a saliency map [9] (as shown in Figure 4). Finally, this map is segmented into three regions of stimuli saliency - the attentional focus. The algorithm here proposed is an extension of the space variant attentional system described in [3] for a stationary robot.

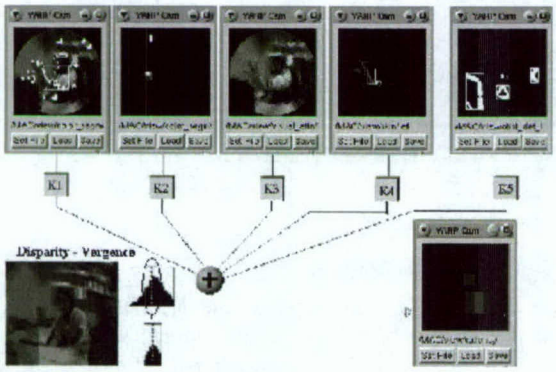

Figure 4: *Although the real world does not avail precise or singularly adequate perceptual information, unique interpretations of the world can be constructed using an attentional system. The basic feature maps are weighted and summed, yielding the saliency map. The latter also receives inputs from the inhibition of return and context-priming processes. The focus of attention is obtained from segmenting this map.*

Logpolar vision - The log-polar transformation maps a cartesian geometry to a non-uniform resolution geometry. We used space-variant images to implement the basic feature detectors.

Color Processing - it includes [3] *i)* a general-purpose color segmentation algorithm based on histograms *ii)* a blob detector, based on region growing.

Skin Detection - based on the algorithm described in [2].

Optical Flow - optical flow is the apparent motion of image brightness [4]. The optical flow constraint assumes that *i)* brightness I(x,y,t) smoothly depends on coordinates (x, y) on most of the image; *ii)* brightness at every point of an object does not change in time, and *iii)* higher order terms are discarded.

The optical flow is computed through a system of linear equations, given by the optical flow constraint for a group of pixels. We apply a coordinate transformation to take into account the logpolar geometry.

Edge Detection - it includes *1)* Gaussian filtering *2)* Sobel edge detector *3)* selection of the five image regions with stronger edges, using a blob detector.

Multiple Levels of Resolution - task dependent selection of the image resolution. Navigation behaviors activate a low resolution mode (visual attention input from wide-field of view cameras), while social behaviors activate high resolution, foveal input.

Context-priming - it includes *i)* modulation of the attentional gains, carried out by the motivational drives and behavioral system, and *ii)* an additional input feature map, that contains context-salient features, such as direction of face gazing.

Inhibition of return - implemented using an inhibition feature map. The map is decremented over the regions of the last detected target, and incremented on a neighborhood of it to model persistence.

Once the attentional system selects regions of the visual field that are behaviorally relevant, more intensive computational resources are applied to these target regions.

3.2 Inertial Perception

The robotic head is equipped with an Intersense gyroscope that estimates the yaw, pitch and roll rotational velocities. This information is used for eye control, for ground slope inference and for body motion detection.

Exponential coordinates were used for rotation. Let Rt be an element of the special orthogonal group SO(3) (see [6]). Then $Rt = e^{\hat{w}\theta}$, where $\hat{w} \in R^3, \|w\| = 1$ belongs to the vector space of all 3×3 skew symmetric matrices, and $\theta \in R$. Let $R_{gyro} = R_{roll} R_{pitch} R_{yaw}$ be the rotation given by the angles from the gyro, $R_{tiltoffset}$ the offset rotational transformation from the gyro to the top of the head, $R_{headtilt}$, $R_{necktilt}$, R_{pan} and $R_{headroll}$ be the transformations due to joint rotations from the base to the top of the head. Then the rotation of the base relative to the ground is given by $Rt = R_{necktilt} R_{pan} R_{headtilt} R_{headroll} R_{tiltoffset} R_{gyro}$.

Applying Rodriguez' formula, the rotation vector $\hat{w}$ is recovered by equation 1:

$$\hat{w} = \frac{1}{2sin(\theta)} \begin{bmatrix} Rt_{32} - Rt_{23} \\ Rt_{13} - Rt_{31} \\ Rt_{21} - Rt_{12} \end{bmatrix} \quad (1)$$

and the rotation angle is obtained from

$$\theta = acos\left(\frac{trace(Rt) - 1}{2}\right), \ \theta_{cart_i} = \theta * w_i \quad (2)$$

Finally, dangerous slopes are estimated by the L2 norm in roll and tilt - $\| [\theta_{cart_1} \ \theta_{cart_2} \ 0] \|$. Platform movement is detected from the mobile platform encoders if available, or from the L2 norm of the base rotational velocities of the robotic head.

3.3 Thermal Vision Perception

Infrared vision is insensitive to light variations, and thus adds extra capabilities to a robot, such as night vision information. It also facilitates people detection.

The thermal image is segmented using a dynamic threshold, and then clustered into regions using a region growing algorithm. A reactive collision detector is implemented by measuring the area's increasing rate of an image region. High increasing rates activates proportionally the self-preservation motivational drive.

4 Architecture

The brain for the M2-M4 Macaco robotic head consists of a flexible, modular and highly interconnected architecture that integrates social interaction, object analysis and functional navigation modules. The software architecture includes, besides the Visual Attention system, releasers from body sensors and motivational drives. Action is determined by competing behaviors, which also share resources to achieve multi-behavior tasking.

4.1 Behaviors and Motivation

Four state variables of the architecture are observed: curiosity, social interaction preferences, navigation and an instinct for self-preservation. These state variables, besides allowing for a richer and more complex selection of behavioral patterns, also provide additional knowledge of the architecture's internal state. Although more variables could be added to the observer, these would impose new design constraints, so that the dynamics from the sensors to the observer and from the latter to the control outputs and feedback connections follow a desired pattern of activation.

Memory was added to the motivational drives, which also receive input from the pre-attentive basic feature detectors. The weighted sum of these values (weighted differently for each drive), is saturated by a sigmoid function. The use of smooth functions was motivated by the existence of well-developed tools to analyze smooth high-dimensional manifolds. The outputs of the motivation drives have feedback connections to the attentional system in order to alter gains of the latter. For instance, a strong activation of a social drive (see also the method described in [2]) will also increase the gains of the social-oriented basic features. The motivational drive with the largest activation pre-selects a set of behaviors, which is modulated by the second most activated drive (if its output is above a threshold). The set of behaviors that best satiates the robot is then selected, as shown in Figure 5.

A low-level luminosity detector triggers reactively the heat detection and the night navigation algorithms, while all other implemented behaviors are inhibited. Despite the competition among behaviors, they also share resources, leading to the emergence of a large spectrum of external behaviors.

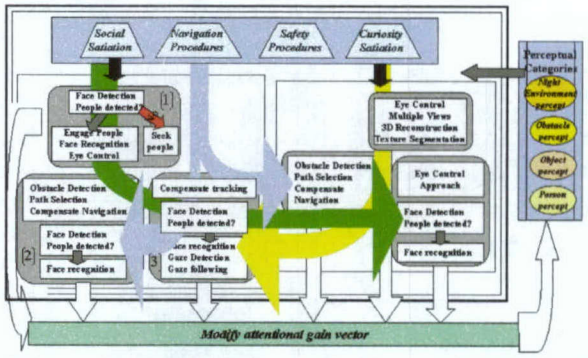

Figure 5: *Diagram for motivations and behaviors. Only social and curiosity behaviors are shown. For instance, assume the social drive gets satiated. If navigation procedures are also activated, then all processes in (2) are enabled. If instead curiosity needs satiation, then (3) is enabled, otherwise (1) is enabled.*

4.2 Eye control

The camera model was simplified to a linear gain, and the robotic eyes were equipped with a set of movements made by frontal eyed, foveal animals:

Saccade and Smooth Pursuit - Ballistic movements are executed without visual feedback. A position error under a threshold enables visual feedback to close the loop using velocity control for smooth trajectories.

Vergence - Wide-field of view stereo images are correlated with each other in logpolar coordinates. Different resolution scales are used to decrease processing times. The disparity is fed into a velocity PD controller, which controls both eyes with opposite signs.

Vestibulo-ocular-reflex (VOR) - Two rotational velocities (yaw, pitch) from the gyro are subtracted to the eyes' velocity command to compensate for body movements. The roll position signal is compensated by a PID position controller, which maintains the eyes baseline parallel to the ground.

5 Object Analysis

This section describes the post-attentional image processing for the extraction of relevant object features.

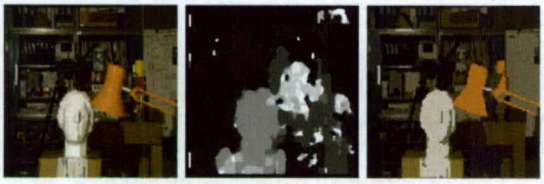

Figure 6: *(left) Left image of Tsukuba stereo images (center) Disparity map (right) Color Segmentation.*

5.1 Texture/Color Segmentation

A Wavelet transform is initially applied to the image. This is equivalent to a family of Gabor filters sampling the frequency domain in a log-polar manner. We correlated the original image with a Daubechies-4 filter using the Mallat pyramid algorithm, at two levels (N=2) of resolution [8], resulting $n = \sum_{i=1}^{N} 4^i$ coefficient images. All these coefficient images were then up-sampled to the original size, using a 5×5 gaussian window for interpolation. This way, more filtering is applied to the coefficients at the N^{th} level. For each pixel, the observation vector is then given by the n wavelet coefficients ($3n$ coefficients for RGB images).

A mixture of gaussians is applied to probabilistically model the input data by clusters. It is therefore necessary to learn the parameters for each cluster and the number of clusters. The former is estimated using the expectation-maximization (EM) algorithm. The latter uses an agglomerative clustering approach based on the Rissanen order identification criteria (MDL), [7]. The image is thereafter segmented according to the cluster of gaussians.

5.2 Matching and Object Reconstruction

A non-linear calibration algorithm is used to automatically calibrate the four cameras. The stereo algorithm first computes the epipolar lines, and then rectifies both images so that the epipolar lines become parallel. The correspondence process is thereafter applied along the same horizontal line. Real-time normalized correlation was selected for matching (other feature-based methods, such as dynamic programming or relaxation techniques [1], do not produce dense maps).

A linear algorithm for minimizing the square error extracts 3D information from matching. From each camera i, with (u_i, v_i) being the image coordinates, M the 3-dimensional vector, and $P_i = [q_{i1}^T \;\; q_{i14};\;\; q_{i2}^T \;\; q_{i24};\;\; q_{i3}^T \;\; q_{i34}]$ the calibration matrix:

$$\begin{array}{l} (q_{i1}^T - u_i q_{i3}^T)M = u_i q_{i34} - q_{i14} \\ (q_{i2}^T - u_i q_{i3}^T)M = v_i q_{i34} - q_{i24} \end{array} \quad (3)$$

The two cameras give 4 equations in the three unknowns M, solved by the pseudo-inverse method.

Curiosity behaviors (see Figure 5) trigger motor movements for multiple view object reconstruction.

6 Social Mechanisms

For the robot to achieve a convincing social role, the vision system is equipped with face detection and recognition modules, together with an algorithm for the detection of human gaze direction.

6.1 Face Detection and Recognition

We use a computationally inexpensive and efficient algorithm (developed by Viola's group at MIT) which locates faces in cluttered scenes. If a face is detected, the algorithm estimates a window containing that face, as shown in Figure 7-left. The cropped image is then fed to the face recognition and gaze inference algorithms.

Figure 7: *(Left) Face detected and window used for recognition. (Right) Heat detection from the thermal camera image.*

We implemented a face recognition scheme that uses embedded Hidden Markov Models (HMM). An embedded HMM is a generalization of a 1D HMM, with a structure similar to a 2D HMM. It consists of a set of superstates where each superstate is an HMM (see [5] for details). As observation vectors we use the energy of the first six coefficients of the Windowed Fourier Transform on local patches of the image. The input image is segmented into five equally spaced vertical regions. The upper vertical region, corresponding to the forehead and hair, is segmented into 4 horizontal regions, and the lower into 3 regions. The other three vertical regions are segmented into seven equally spaced regions, since these segments correspond to the most important features of a face. Hence, the size of the observation vector is $6 \times (4 + 7 \times 3 + 3) = 168$. We achieved real-time face recognition after of-line training with a database of six people.

6.2 Head Gaze Direction

A supervised learning strategy was adopted, being the classification problem separated into three classes: left, center and right gaze (although its generalization to nine classes is straightforward). A mixture of gaussians was used to model each class, as was done for the single class of the texture segmentation scheme.

The original image is initially convolved with a wavelet at two levels of resolution. The coefficients on the second level are averaged to achieve a spatial resolution of 16×16, for all images on the training set. Afterwards, the dimensionality of the input vector ($16 \times 16 \times 4^{N=2} = 4096$) is reduced by applying Principal Component Analysis.

Classification is performed by computing the spectral vector for an image, and then determining the a posteriori probability that this vector belongs to a class, selecting the maximum corresponding class.

7 Navigational Mechanisms

The inputs of the navigation algorithm are stereo clusters of strong image edges detected by the pre-attentional mechanism, as shown for one image in Figure 8. Stereo points belonging to these clusters are matched using a dynamic programing algorithm [1]. Three-dimensional information is then extracted from the feasible matches, and clustered into five groups using k-means. These groups are interpreted as obstacles - objects, people or walls.

Figure 8: *(left) Blobs of strong edges detected from (right) a rectified image*

7.1 Night Navigation

Although 3D information is lost with low visibility, the platform is able to operate thanks to a thermal camera. A navigation algorithm based on monocular cues runs at frame-rate, for night navigation. In addition, regions of the image that are potential candidates as sources of heat, like living organisms, are segmented from the image, together with a measure of the approximation of these sources relative to the robot.

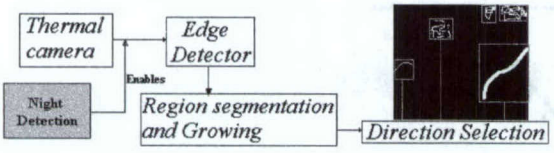

Figure 9: *Night navigation algorithm*

8 Conclusions and Future Work

We presented a project of a portable robotic head, flexible at the mechanical, hardware and software levels. Although equipped with a complex cognitive system, its small weight and compact size allows it to be incorporated into multiple mobile platforms, and also to be used as a portable autonomous sociable robotic creature.

We are addressing the extension of the gaze inference algorithm to account quantitative measures of head rotation. We are also currently working on object and scene recognition from contextual cues,for future application in both dynamic objects identification and mobile robot localization.

Acknowledgments

This work is funded by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Clues" project under contract number DABT 63-00-C-10102. The author was supported by a Portuguese government grant PRAXIS XXI BD/15851/98.

References

[1] A. Arsenio and J. Marques, "Performance Analysis and Characterization of Matching Algorithms," *5th Int. Symp. on Intelligent Robotic Systems,* 1997

[2] C. Breazeal, "Sociable Machines: Expressive Social Exchange Between Humans and Robots," *Doctoral Dissertation,* EECS-MIT, 2000

[3] G. Metta, "An Attentional System for a Humanoid Robot Exploiting Space Variant Vision," *IEEE-RAS Int. Conf. on Humanoid Robots,* 2001

[4] B. Horn, *Robot Vision,* MIT Press, 1986

[5] S. Kuo and O. Agazzi, "Keyword spotting in poorly printed documents using 2-D HMMs," *IEEE PAMI,* Vol. 16, pp. 842-848, 1994

[6] Murray, Li and Sastry, *Robotic Manipulation,* CRC Press, 1994

[7] J. Rissanen, "A Universal Prior for Integers and Estimation by Minimum Description Length," *Annals of Statistics,* Vol. 11, pp. 417-431, 1983

[8] G. Strang and T. Nguyen, *Wavelets and Filter Banks,* Wellesley-Cambridge Press, 1996

[9] J. Wolfe, "Guided Search 2.0: A Revised Model of Visual Search," *Psychonomic Bulletin and Review,* n. 1, pp. 202-238, 1994